CSE-6363-003

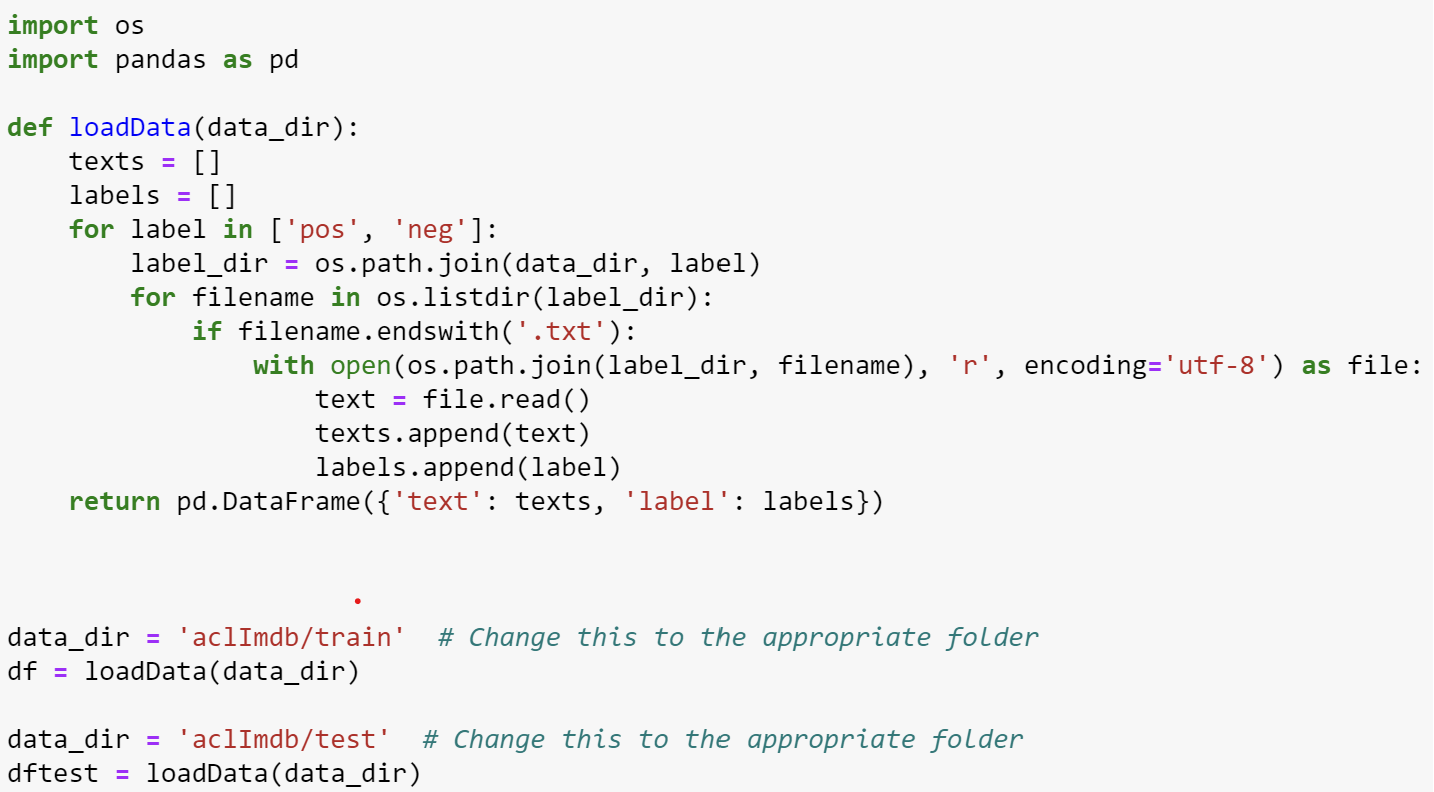
Sentiment Analysis on Movie Review

# **Team 16:**

* Harshita Chegondi - 1002115738
* Pranav Rayapati - 1002077440
* Rakshitha Koushika – 1002076118

**Data Preprocessing:**

In this project, we have pre-processed data before training and testing of data. Firstly, data has been extracted from folder then segregated the texts according to the labels.

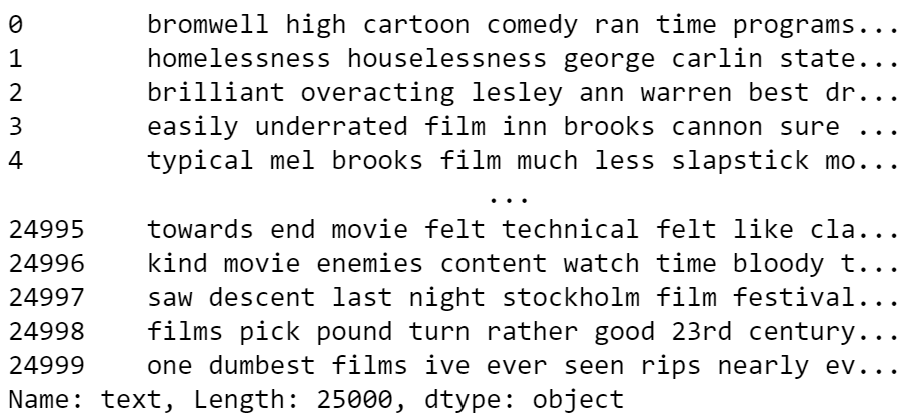


After this, the labels have been encoded and removed stopwords, punctuation, HTML tags then converted to lower case. After filtration, the data looks like this:

A black text on a white background

Description automatically generated

Lemmatization is used to help understand the tone of a sentence. In sentiment analysis, it is mostly used to detect positive and negative sentence. The below snippet shows the data after lemmatization.



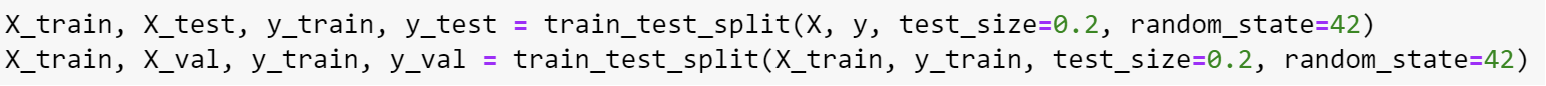
Next, Tokenizer have been applied. Tokenizer is applied because it enables machine learning to understand the large amount of text data.

A white sheet with numbers and a black dot

Description automatically generated with medium confidence

**Splitting the datasets:**

Splitting of the dataset is very important as we get train accuracy and test accuracy from a dataset. In this project we are dividing the dataset to 80:20 where 80% data is used for train data and 20% data is used for test data



**Models used in this project:**

*Convolutional Neural Network:* this model has been used to get the best accuracy from the given dataset. We have used set of layers to implement CNN and passed train data and test data to get the best accuracy. The below snippet shows the implementation of model.

A computer code with text

Description automatically generated

Epochs for train data from CNN:

A white background with black text

Description automatically generated with medium confidence

Accuracy from validation data:

A screenshot of a computer code

Description automatically generated

Graphical Representation of Model loss and Model Accuracy from CNN

A graph of a model loss

Description automatically generated *A graph with a line graph and a line graph

Description automatically generated*

From the graphs, there is no overfitting from the model and hence this is the best fit with low loss and and high accuracy

*Bidirectional Encoder Representations from Transformers(BERT):* this model is used to understand and handle ambiguous texts and predicts the sentiment from it. Below is the BERT model.

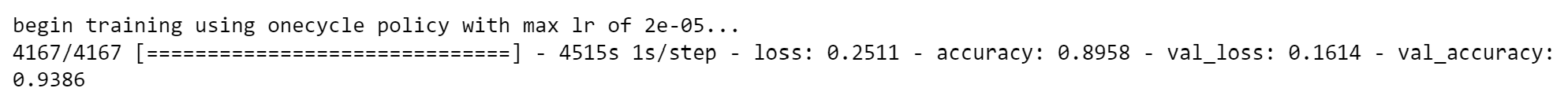
A close-up of a sign

Description automatically generated

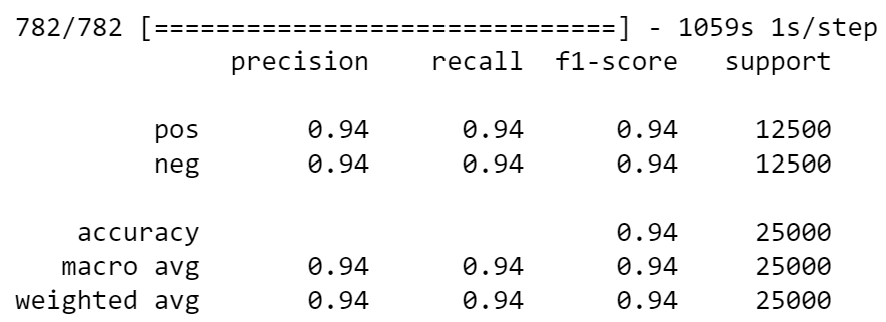
A black text on a white background

Description automatically generated

Epochs for train data from BERT:



Accuracy from validation data in BERT model:



Comparisons of accuracies from both the models:

|  |  |
| --- | --- |
| Accuracy of CNN | Acccuracy of BERT |
| 90% | 94% |

As per the observation from both the accuracies, BERT is the best model when compared to CNN because BERT model has given good accuracy. Both both the models are the best as per the usage.

References:

* A Hybrid BiLSTM-ATT Model for Improved Accuracy Sentiment Analysis - Langxue Dang, Chunyu Wang, Hongyu Han, Yan-E Hou – 2022

<https://ieeexplore-ieee-org.ezproxy.uta.edu/document/10074634>

* Comparative Study of Algorithms for Sentiment Analysis on IMDB Movie Reviews – Neelisetty Sri Lakshmi Sai Charitha, Vellanki Rakesh, Mandadapu Varun – 2023

<https://ieeexplore-ieee-org.ezproxy.uta.edu/document/10113113>

* Sentiment Prediction of IMDb Movie Reviews Using CNN-LSTM Approach – Mahesh Mishra, Amol Patil – 2023

<https://ieeexplore-ieee-org.ezproxy.uta.edu/document/10165155>

* ktrain: A Lightweight Wrapper for Keras to Help Train Neural Network

<https://towardsdatascience.com/ktrain-a-lightweight-wrapper-for-keras-to-help-train-neural-networks-82851ba889c>

* BERT Text Classification in 3 Lines of Code Using Keras

<https://towardsdatascience.com/bert-text-classification-in-3-lines-of-code-using-keras-264db7e7a358>